Sampling Strategies in a Statistical Approach to Clinical Classification

Yiming Yang, Christopher G. Chute
Section of Medical Information Resources, Mayo Clinic/Foundation
Rochester, Minnesota 55905 USA

This paper studies the sampling strategies for the Expert Network (EexNet), a statistical learning system used for patient record classification at the Mayo Clinic. The goal is to achieve high accuracy classification at an affordable computational cost in very large applications. The learning curves of ExpNet were observed with respect to the choice of training resources, the size, vocabulary coverage and category coverage of a training set, and the category distribution over training instances. A method combining advantages of different sampling strategies is proposed and evaluated using a large training corpus. As a result, Expert Network has achieved its nearly-optimal classification accuracy (measured by average precision) using a relatively small training set, with a fast real-time response which satisfies the needs of human-machine interaction.

INTRODUCTION

Assigning predefined categories to free texts (*text classification* or *text categorization*) has wide application since categories are often used to index real-world databases. At the Mayo Clinic, for example, about 1.6 million diagnoses in patient records are coded annually using HICDA-2 (the Hospital Adaptation of ICDA, 2nd Edition)[1] for the purposes of billing and research. Manual categorization remains the dominant method in practice, which is both costly and error-prone. To improve the quality and to reduce the cost, we have developed several classification systems and used them to assist human classification in the Section of Medical Information Resources, Mayo Clinic. A preliminary evaluation[2] showed that both classification accuracy and coding speed were improved by using these tools, and that the systems which use statistical learning techniques are more effective than the systems which search categories based on string matching between texts and category names or definition phrases.

Our statistical classification systems are named the Linear Least Squares Fit (LLSF) mapping and the Expert Network (ExpNet). LLSF is a regression method which predicts the categories of a new text based on the correlations between the words and categories of training texts [3]. ExpNet is a Nearest Neighbor (NN) classification method which ranks candidate categories for a new text based on the categories of its neighbors in training texts[4] [5]. The two methods were almost equally effective with respect to classification accuracy in our experiments when using the same training data. Both significantly outperformed alternative methods such as word-based matching methods which do not use any human knowledge, and thesaurus-based or rule-based methods which are heavily dependent on manually coded human knowledge. Other studies [6] showed superior results of NN classification over a rule-based approach using manually coded expert knowledge. Comparison of LLSF and ExpNet with more sophisticated learning methods such as neural networks, Bayesian belief networks and classification-tree methods remains an open area of research. Part of the difficulty is that none of the more sophisticated methods easily scale to very large text classification problems with thousands or tens of thousands of categories, due to computational tractability issues. LLSF and ExpNet are relatively efficient and therefore scale more easily. Solving a linear regression is generally simpler than finding a non-linear solution using neural networks, and the computation in NN classification is even more efficient[4]. Both LLSF and ExpNet have their strengths and weaknesses. LLSF requires an intensive off-line training, which makes it more difficult than ExpNet to scale up. However, once the training is done, the on-line category ranking for a given text is relatively fast. ExpNet needs little training in advance, but requires an on-line search through unique training texts [4] for the NNs of each testing text[1]. Hence, attaining real-time response for the NN search is a computational problem when the training set is very large. Some techniques for scaling-up LLSF were reported in a separate paper [7]. Here we focus on ExpNet and its application to a very large and practical problem, the classification of diagnoses in Mayo patient records using HICDA-2 categories.

The research interest in this paper is on sampling strategies for statistical learning given a method. We address the question of how to obtain a training set which contains sufficient information and is of a reasonable size. We choose ExpNet over LLSF for this study because it allows us to explore a large problem which cannot be handled by the current LLSF system. There are 29,741 categories in HICDA-2, and about 2.4 million diagnoses (DXs) coded by humans with computer assistance in our section each year. These coded DXs are all eligible for training in ExpNet. However, our current production system has only used a subset of 205,660 DXs (a few months

[1]ExpNet uses a vector of word weights to represent a text, and the cosine value of two vectors to measure the similarity of two texts. The cosine value reflects how much the two texts are in common, in terms of shared words.

0195-4210/95/$5.00

accumulation) plus the 29,741 HICDA-2 definition phrases as the training set, for which the ExpNet has an on-line response time of about 1.5 seconds per DX when using a SPARCstation 10. We cannot use all available DXs for training because this would make the on-line response time too slow for the user-machine interaction. For example, if we use 5 million DXs for training, the response time would increase to 32 seconds per DX; if we use 10 million DXs, then the response time would be more than one minute. The point is, we must decide what to do with the large amounts of available training data. Potential solutions for a large NN classification problem include:

- hardware solutions, e.g. Thinking Machine's work using massive parallel computers to speed up the NN search [6] [8];
- software solutions, e.g. to partition training instances in advance and to search selectively instead of exhaustively [9]; or
- sampling solutions, i.e. to find a relatively small training set which has sufficient information for the classification.

This paper addresses the third direction. We want to know how large a training set is necessary before pursuing hardware or software solutions for a very large problem. Statistical sampling theories have shown that one can obtain a fairly accurate estimate about a large population using a relatively small sample. The question here is, how do we minimize a training set for ExpNet without losing useful information for text classification.

SAMPLING STRATEGIES

Several questions need to be answered when choosing sampling strategies:

- What kind of texts should be used for training? Should we use coded DXs only, or should we use HICDA-2 category definition phrases in addition?
- What criteria should be used to judge a training set? Which criterion is more important, the usefulness (how effective it is in terms of categorization accuracy and computational efficiency) or the completeness (the coverage of possible DXs)?
- How do we measure the usefulness and completeness of a training set? Are vocabulary size and coverage of unique categories important measures? Is DX instance distribution over categories more important than vocabulary coverage and category coverage?

We believe that Mayo DXs should be the major resource of training data, because instances from the application itself represent the application the best. The category definition phrases should be used only if there is experimental support of their usefulness.

We prefer usefulness over completeness. The usefulness of a training set should be judged based on its effects on ExpNet in classifying the majority of DXs, not just a few cases. Computational efficiency should also be counted in measuring the usefulness.

We think that DX distribution over categories in a training set would be more informative than vocabulary coverage and category coverage for analyzing the impact of the training set on the effectiveness of a statistical classification method. Not every category is equally important for training. For example, the training collection of 205,660 DXs mentioned before contains 234,465 category instances in which 66% of the HICDA-2 categories are missing, 10% of the total categories had only one instance or 3027 instances together, and 1.4% of the total categories has more than one hundred instances each or 146,964 instances together. This means that if ExpNet fails to classify the 76% rare categories which have one instance or less, the expected error rate is 3027/234465 or 1.2%. On the other hand, if ExpNet fails to classify the 1.4% most common categories, then the expected error rate is 146964/234465 or 62.7%. Clearly, whether common categories are well represented in a training set is crucial for the global effectiveness of classification, while missing a large number of rare categories in a training set may have a statistically insignificant impact only. Since the NN classification follows a *majority-vote* principle, it often favors the categories with more instances in the training set, than those which are not. To optimize the global effectiveness, we want the more important categories (as measured by frequency) to be better represented than the less important ones in the training sample. Pursuing an even coverage for all categories or words, therefore, is not a suitable sampling strategy.

To verify our assertions, we included the following sampling strategies in this study:

1) *Natural-occurrence sampling*: leave the DXs in the order of the time they were collected from patient records and coded by humans, and take a continuous chunk as a training set. This is the current strategy of our production system (although we have only used the first chunk of accumulated DXs). The advantage of such a sampling method is that a training set would naturally reflect the category distribution in the population, and that the classification of ExpNet would favor common DXs over rare ones. A potential weakness is that a training set may have too many redundant instances of common categories, and may not have enough coverage of rare categories.

2) *Completeness-oriented sampling*: prefer some instances over others if they contribute more new words or categories to a training set. An extreme example of this would be to take the HICDA-2 definition phrases only as a training set or as the dominating part of a training set, because they contain all the

unique categories, and more unique words than a natural chunk of DXs would contain. Such a method ignores the natural distribution of categories in the original population, and may consequently decrease the over-all classification effectiveness of ExpNet because the common cases are not well represented in a training set. A potential advantage of such a strategy is to have better coverage of rare cases.

3) *Sort-and-split sampling*: sort the DXs by categories first, and sort the DXs with the same category in alphabetical order, then split the sorted DXs into k subsets in a way that the first DX in each k DXs belongs to the first subset, and the second DX in each k DXs belongs to the second subset, and so on. The attempt is to combine the advantages of the two strategies mentioned above, that is, to favor the majority cases and also to have reasonable coverage of rare cases. For example, when applying such a strategy with $k = 2$, every category with two or more instances in the population will have at least one instance in a training set. On the other hand, a category with 100 instances will have 50 instances in the training set. The absolute difference is reduced, but the relative difference remains.

EMPIRICAL VALIDATION

The Data

A collection of eligible training data is needed as a pool from which different sampling strategies can be applied. The training set of our current production system is chosen for such a purpose, which consists of 205,660 diagnoses from Mayo medical records in the period of October 1993 to March 1994, and the 29,741 category definition phrases in HICDA-2. A diagnosis is a descriptive free-text with 1-26 words, or 3 words per DX on average. About 88% of these diagnoses have a uniquely matched category; the rest have 2-7 categories. A category definition in HICDA-2 has 3 or 4 words on average. For convenience, we use *text* to refer to either a DX text or a category definition phrase, and *code* to mean the unique identifier of a category. The DXs and the category definition phrases together form a collection (*the superset*) of 235,401 training texts each of which has one or more category codes assigned by humans or as defined in HICDA-2. The vocabulary size of these training texts is 15,994 unique words. Subsets are derived from this superset, according to different sampling strategies.

A testing set is selected for evaluating the sampling strategies. There are five testing sets which were collected for our previous evaluation of different classification methods [2]. Each set consists of about 1000 DXs arbitrarily chosen from the patient records at the time of that evaluation. By checking the common words and categories of each of these testing sets and the training superset mentioned above, we found that these testing sets are similar in the sense that they all have about 97-98% of the words and 96-97% of the categories covered by the training DXs mentioned above. Hence, one of the five testing sets was arbitrarily chosen for this study, containing 1071 DXs, 1249 unique words and 726 unique categories.

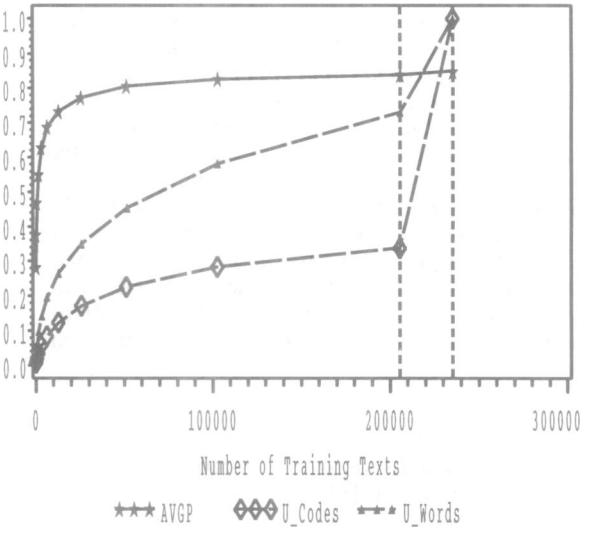

Figure 1: Learning curve of ExpNet using sort-and-split sampling.

Preprocessing was applied to these training and testing sets for the removal of punctuation and numbers, and for changing uppercase letters to lowercase; no stemming or removal of "noise words" was applied.

The Results

Figure 1 shows the learning curve of ExpNet in response to the number of texts or *the size* of a training subset. The training sets are derived using the sort-and-split sampling strategy mentioned in the previous section. The 205,660 DXs were sorted by codes first, and the DXs with a same code were sorted in alphabetical order. A subset was obtained by selecting the first of every i DXs. By setting split parameter k to 2, 4, 8, ..., 1024, subsets with sizes of 200, 401, 803, ..., 102,830 DXs were obtained. These subsets were used as training sets for ExpNet, and evaluated using the testing set of DXs mentioned before. The full set of 205,660 DXs and the superset of 235,401 texts were also evaluated. To evaluate the classification effectiveness, we computed the conventional ten-point average precision (AVGP) of category ranking, i.e. we computed the precision values at recall thresholds of 10%, 20%, ... 100%[10], and averaged these values as a global measure. The results are interpolated into the star-curve in Figure 1; we call it *the learning curve* of ExpNet. The dashed lines correspond to the results of using the full set (205,660 Mayo DXs) and the superset (205,660 Mayo DXs plus 29,741 HICDA-2 definition phrases), respectively. The triangle-curve shows the ratio of the number of unique words in a training set divided by the total number of unique words in the superset. The diamond-curve shows the ratio of the number of unique categories in a training set divided by the total number of unique categories in the superset. The

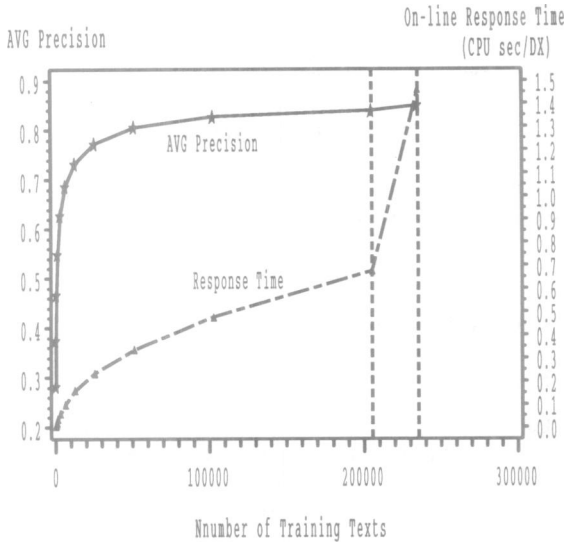

Figure 2: Trade-off between effectiveness and efficiency of ExpNet.

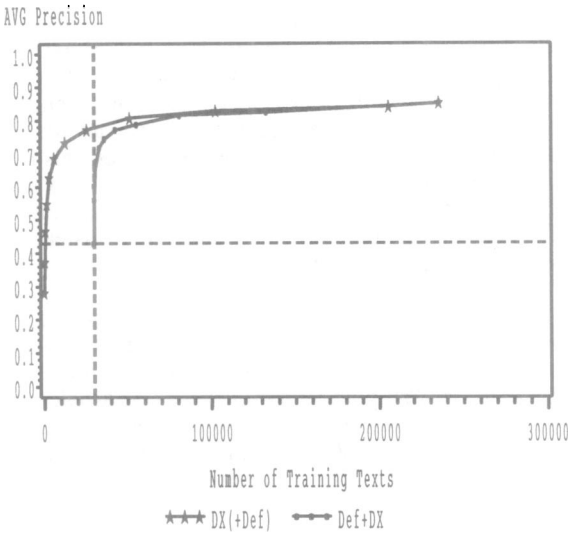

Figure 3: Learning curves of ExpNet using different sampling strategies.

interesting points are:

1) The learning curve rises rapidly when the training sets are small, and becomes relatively flat when the training set size achieves 100,000 texts or larger. This means that high-frequency DXs were included even in a small training set, and that these DXs were more influential in the over-all performance than rare DXs. This learning curve also indicates that further increase in the size of a training set beyond the 200,000 level is unlikely to have significant improvement.

2) The slope of the unique-word curve and the unique-category curve is much larger than the learning curve for most of the regions, except the very left end of this graph. This means that the improvement in word coverage and category coverage of a training set does not necessarily transfer into an improvement in classification effectiveness. In other words, a large number of words and categories are not crucial for the global classification performance because they are rare.

3) Adding the HICDA-2 phrases to the training DXs did not improve the AVGP by much, although it significantly increased the vocabulary and categories coverage of the training set. This means that most of the words and categories which are needed for classification are already included in the training DXs, and that the category definition phrases contribute little useful information to the training.

Figure 2 shows the trade-off between AVGP and the on-line response time of ExpNet. The average CPU seconds for category ranking per DX was measured. A significant time increase was observed when the HICDA-2 definition phrases are added to the training set, because most of these definition phrases are unique to the entire collection, and the response time of ExpNet is proportional to the number of unique training texts. Clearly, using HICDA-2 in addition to Mayo DXs for training doubled the computation time for only an insignificant improvement in classification effectiveness.

Figure 3 compares the learning curve in Figure 1 with the learning curve (dot-curve) when the training sets were constructed in a different way. That is, the HICDA-2 phrases were used as the basic training set, and each of the subsets of the Mayo DXs was added to the basic set respectively. When using the 29,741 HICDA-2 phrases alone for training, the AVGP was only 42% which is between the results of using 400-800 DXs for training, and is similar to the performance level as applying a word-based matching between DXs and HICDA-2 phrases (which had an AVGP value of 44% in our experiment). On the other hand, when using a similar amounts (25,707) of DXs instead of the 29,741 HICDA-2 phrases for training, the AVGP was 77%, or a 83% relative improvement over the result of using the HICDA-2 phrases. Comparing the two curves in this graph, it is clear that given any fixed size of a training set, the use of HICDA-2 phrases plus DXs had no significant improvement over using DXs alone, if any. However, the on-line computation time in the former is much higher than in the latter, because the HICDA-2 phrases contribute more unique texts to a training set than DXs do, as we discussed before.

The natural-occurrence sampling was also tested. Leaving the 205,660 DXs in the order of time when they were collected from patient records, we divided them into four equally-sized subsets sequentially. Each of these subsets was used for training; the AVGP values was 77.6% on average. For comparison, we also tested four training subsets with the same size but

using the sort-and-split strategy; the averaged AVGP value was 80.0%. The improvement of the latter over the former comes from a better balance between high-frequency and low-frequency categories. That is, the influences of high-frequency categories and low-frequency categories are adjusted in a way that globally improved the results. This experiment also indicates that applying the sort-and-split strategy to a larger pool of training data, to draw a subset with the same size as the current superset, and use it instead of the superset, we may have an improved result.

Note that in all the above experiments, we did not apply removal of "noise words" because this is not the focus of this paper. When applying word removal using a standard "stoplist" which consists of articles, prepositions, conjunctions and some common words, the AVGP on the 205,660 DX training set was improved from 84% to 86%; the on-line response of ExpNet was 0.7 seconds per DX.

CONCLUSIONS

This paper studied the sampling strategies for ExpNet in its application to clinical classification at the Mayo Clinic. The major findings include:

1) The use of manually coded DXs for training is a better choice than using both the DXs and the HICDA-2 category definition phrases, because the latter contributes little useful information at a high computational cost. Using the sort-and-split sampling strategy to select a training set from a large DX collection is better than using *a natural chunk* of DXs, given a fixed size of the training sets.

2) ExpNet has achieved its nearly-optimal classification performance when using a training set of 205,660 DXs. The 86% precision on average with an on-line response of 0.7 second per diagnosis is highly satisfactory for the current needs in our computer-assisted classification applications. The learning curve of ExpNet indicates that significant improvement beyond this point would be difficult, even if the size of the training set is significantly increased. Using unnecessarily large training data would substantially decrease the efficiency of the system, on the other hand.

Finally, no claim is made that the particular size of an optimal or nearly-optimal training set for one application is generalizable for all applications. The optimal training set size for Mayo diagnosis classification may not be the optimal size for MEDLINE document classification, for example. Given that a diagnosis phrase has three words on average, and that a MEDLINE article has typically 150-200 words in its title and abstract, the necessary amounts of training data may be larger for the latter than for the former. Nevertheless, the analysis method presented here is generalizable to other domains/applications and alternative statistical classification methods. Future research topics include a sampling strategies analysis for ExpNet in categorization of Mayo patient-record data other than diagnoses and MEDLINE documents, and similar analyses for the Linear Least Squares Fit mapping method in MEDLINE document indexing and Mayo patient record classification.

ACKNOWLEDGEMENT

This work is supported at Mayo Foundation in part by NIH Research Grants LM05714 and LM05416.

References

[1] *HICDA-2, Hospital Adaptation of ICDA, 2nd Edition.* Ann Arbor, MI: Commission on Professional and Hospital Activities, 1968.

[2] Chute CG, Yang Y. Buntrock J. (1994) An evaluation of computer-assisted clinical classification algorithms. *18th Ann Symp Comp Applic Med Care (SCAMC) JAMIA 1994;18(Symp.Suppl)*:162-6.

[3] Yang Y, Chute CG. (1994) An example-based mapping method for text categorization and retrieval. *ACM Transaction on Information Systems (TOIS 94)*: 253-277.

[4] Yang Y. (1994) Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval. *17th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 94)*: 11-21.

[5] Yang Y, Chute CG. (1994) An application of Expert Network to Clinical Classification and MEDLINE indexing. *18th Ann Symp Comp Applic Med Care (SCAMC 94) JAMIA 1994;18(Symp.Suppl)*:157-61.

[6] Creecy RH, Masand BM, Smith SJ, Waltz DL (1992). Trading MIPS and memory for knowledge engineering: classifying census returns on the Connection Machine. *Comm. ACM*, 35, 48-63.

[7] Yang Y. (1995) Noise Reduction in a Statistical Approach to Text Categorization. *18th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 95)*: to appear.

[8] Masand B., Linoff G., Waltz D. (1992) Classifying News Stories using Memory Based Reasoning. *15th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 92)*: 59-64.

[9] Deng K, Moore AW. (1995) Multiresolution Instance-based Learning. *Proceedings of Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)*: to appear.

[10] Salton G. (1989) *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Reading, Pennsylvania: Addison-Wesley.